PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :		(11) International Publication Number:	WO 92/01933
G01N 33/00	A1	(43) International Publication Date:	6 February 1992 (06.02.92)

(21) International Application Number: PCT/US91/01106

(22) International Filing Date: 27 February 1991 (27.02.91)

(30) Priority data:
556,239 20 July 1990 (20.07.90) US

(71) Applicant: E.I. DU PONT DE NEMOURS AND COMPANY [US/US]; 1007 Market Street, Wilmington, DE 19898 (US).

(72) Inventors: SALEMME, Francis, Raymond ; 107 Marshall Bridge Road, Kennett Square, PA 19348 (US). WENDOLOSKI, John, Joseph ; 1418 Jan Drive, Wilmington, DE 19803 (US).

(74) Agents: MEDWICK, George, M. et al.; E.I. du Pont de Nemours and Company, Legal/Patent Records Center, 1007 Market Street, Wilmington, DE 19898 (US).

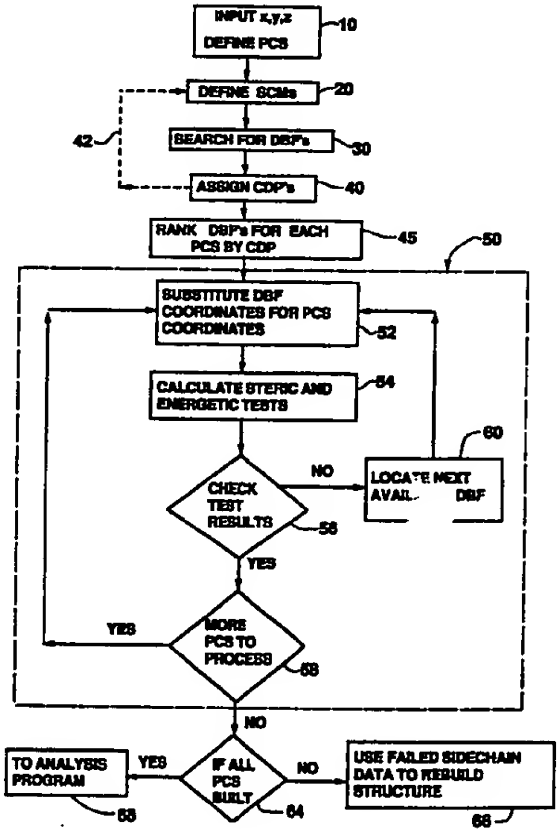
(81) Designated States: AT (European patent), BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), NL (European patent), SE (European patent).

Published
With international search report.

(54) Title: GENERATION OF A COMPLETE SET OF STRUCTURAL COORDINATES OF A MOLECULE FROM A SET OF PARTIAL COORDINATES

(57) Abstract

A complete macromolecular structure (M) is constructed based upon two or more partial coordinate strings (W) representing the three-dimensional coordinates of the location of atoms (A) known to reside in some predetermined sequence at some predetermined region of the molecule. A library of known protein structures is searched to identify each structural fragment that has at least some predetermined subset of the atoms thereof with coordinates that correspond within predetermined limits imposed by a predetermined structural context mask to the coordinates of the atoms in each partial coordinate string (W). Each fragment identified by the search has assigned to it a context dependent probability quantifying the probability that the particular structural fragment corresponds to the partial coordinate string (W). Once all of the identified database fragments have been ranked in accordance with their context dependent probabilities, the coordinates of the atoms forming the structural fragment having the highest context dependent probability is substituted into the partial coordinate string (W), thereby defining an updated partial coordinate string (W). The resultant updated partial coordinate string (W) is tested in accordance with predetermined test parameters.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	ES	Spain	MG	Madagascar
AU	Australia	FI	Finland	ML	Mali
BB	Barbados	FR	France	MN	Mongolia
BE	Belgium	GA	Gabon	MR	Mauritania
BF	Burkina Faso	GB	United Kingdom	MW	Malawi
BG	Bulgaria	GN	Guinea	NL	Netherlands
BJ	Benin	GR	Greece	NO	Norway
BR	Brazil	HU	Hungary	PL	Poland
CA	Canada	IT	Italy	RO	Romania
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland			SN	Senegal
CI	Côte d'Ivoire	KR	Republic of Korea	SU+	Soviet Union
CM	Cameroon	LI	Liechtenstein	TD	Chad
CS	Czechoslovakia	LK	Sri Lanka	TG	Togo
DE	Germany	LU	Luxembourg	US	United States of America
DK	Denmark	MC	Monaco		

+ **It is not yet known for which States of the former Soviet Union any designation of the Soviet Union has effect.**

GENERATION OF A COMPLETE SET OF STRUCTURAL COORDINATES OF A MOLECULE FROM A SET OF PARTIAL COORDINATES

BACKGROUND OF THE INVENTION

Field of the Invention The present invention relates to the generation of a complete set of structural coordinates of a molecule from a set of partial coordinates of atoms of that molecule, and in particular for predicting the three dimensional shape of a molecule, such as a protein molecule.

Description of the Prior Art Protein engineering and rational drug design technologies rely on a detailed structural knowledge of protein structures at the atomic level, Hol, "Applying Knowledge of Protein Structure and Function", (1987) Tibtech, Volume 5, 137-143. Such information is directly accessible through experiment using methods of x-ray crystallography and, more recently, Nuclear Magnetic Resonance (NMR) techniques. However, these methods present formidable technical difficulties, so it is useful to have methods that reliably predict protein structural coordinates from amino acid sequence data.

While the prediction of protein structural coordinates is not presently possible in the general case, computer methods

that can make extensions from partial coordinate data derived experimentally, r investigate properties of structure with sequences that do not exist naturally, offer substantial utility in protein engineering and rational drug design. Exemplary of such computer methods are those disclosed in Jones and Thirup, "Using Known Substructures In Protein Model Building and Crystallography", (1986) Embo. J., 5, 819-822 and in Blundell et al., "Knowledge-Based Prediction Of Protein Structures And The Design Of Novel Molecules", (1987) Nature, 326:6111, 347-352.

The protein folding problem (i.e. the the determination of a protein's three-dimensional structure from its polymeric amino acid sequence) is a central problem in biochemistry. Interest in this problem has recently intensified owing to several factors. These include: (1) The advent of rational drug design methods that depend on a detailed knowledge of protein three dimensional structure; (2) The emergence of DNA sequencing and related technologies that allow the rapid isolation and determination of the protein amino acid sequences of enzymes and receptors associated with disease states; and (3) The development of recombinant DNA methods that allow the facile manipulation and generation of proteins with novel amino acid sequences and properties. Taken together, these provide strong motivation to be able to predict protein three dimensional structure from amino acid sequence, or alternatively, to predict structural and functional consequences of substituting one amino acid for another in a given protein sequence.

At present the protein folding problem is not generally tractable at the level of detail required to direct drug design or protein engineering work. It is not currently possible to predict a protein's three-dimensional structure from its amino acid sequence with accuracy that is comparable with experimental

methods. Consequently, substantial efforts have been directed at developing methods that either assist in making the experimental determinations of three-dimensional protein structure more rapid, or rely on a combination of known structural precedents and computational energy methods to derive three dimensional models for structures that are unknown experimentally.

X-ray crystallography is the experimental method in predominant use for the determination of the three dimensional structure of protein molecules. In this method experimental diffraction data from a single crystal is processed to produce an electron density map which represents the ordered contents of a crystal unit cell. Interpretation of this electron density map in terms of a detailed chemical structure (usually inferred from independently obtained information such as the DNA sequence of the gene that encodes the protein) is a difficult and very laborious task, particularly at the initial stages when the density is an incomplete or an imperfect representation of the chemical structure. Usually it is possible to experimentally obtain an electron density map from which the alphaCarbon (also herein, "alphaC") backbone structure of the protein can be approximately located, although there is more ambiguity about the detailed position of other backbone atoms, and particularly, the conformationally flexible amino acid side chains.

Work by Jones and Thirup, "Using Known Substructures In Protein Model Building and Crystallography", (1986) Embo. J., 5, 819-822 showed that fragments derived from a protein structural database can be used in an interactive computer graphics program to aid in fitting protein molecular models to crystallographically produced electron density maps. Starting with an approximate set of alphaC positions in an electron density map, the remaining backbone atoms (N, C, O and

betaCarbon) could be located by substituting in fragments, derived from a database of known protein structures, whose alphaC positions corresponded closely to the approximate ones defined initially. This means that in the majority of cases the backbone alphaC positions are sufficient to determine the conformation and position of the remainder of the polypeptide backbone atoms. Once the backbone has been fit, the remaining structural ambiguities concern the amino acid side chain conformations, which are in principle free to rotate about the molecular single bonds that they contain.

However, surveys of side chain conformation in known protein structures showed that these tended to be restricted to only a few torsional values and that the observed values could be grouped into a rotomer library for most of the amino acids excepting some that are unusually flexible and usually are situated on the protein exterior. Articles by Janin et al., "Conformation Of Amino Acid Side-Chains In Proteins" (1978) J. Mol. Biol., 125, 357-386 and Bhat et al., "An Analysis of Side-Chain Conformations In Protein Structures", (1979) Int. J. Pept. Res., 13, 170-184) discuss side chain conformational analysis. An article by Ponder and Richards, "Tertiary Templates For Proteins", (1987) J. Mol. Biol., 195, 775-791 shows that a restricted set of amino acid side chains, oriented with low energy rotomer conformations, can in principle be used to compactly pack a protein interior.

Copending application Serial Number 07/409,487, filed September 19, 1989 in the names of Steeg and Hinton discloses and claims a symmetrically connected parallel distributed processing network wherein each node in the network represents the potential pairing of bases in the molecule, with selected ones of the nodes being connected to predetermined other ones of the nodes by connection lines each having a predetermined sign and a value lying within a predetermined

range. The sign and value are representative of energy constraints on the conformation of the molecule.

SUMMARY OF THE INVENTION

In general, the present invention builds a complete macromolecular structure based upon two or more partial coordinate strings representing the three-dimensional coordinates of the location of atoms known to reside in some predetermined sequence at some predetermined region of the molecule. The string could contain atoms that are present along the backbone of the molecule, along side chains extending therefrom, or other atoms incorporated into, or bound to, a protein molecule. A library of known protein structures is searched to identify each structural fragment that has at least some predetermined subset of the atoms thereof with coordinates that correspond within predetermined limits imposed by a predetermined structural context mask to the coordinates of the atoms in each partial coordinate string.

Each structural fragment identified by the search has assigned to it a context dependent probability quantifying the probability that the particular structural fragment corresponds to the partial coordinate string. Once all of the identified database fragments have been ranked in accordance with their context dependent probabilities, the coordinates of the atoms forming the structural fragment having the highest context dependent probability is substituted into the corresponding partial coordinate string, thereby defining an updated partial coordinate string. The resultant updated partial coordinate string is tested in accordance with predetermined test parameters. If the test results are satisfactory, according to a predetermined criterion, substitution of the fragment having the next highest probability into its corresponding partial

coordinate strings in the molecule is made and the resultant updated partial coordinate string is tested.

If the test result for a given updated partial coordinate string is not satisfactory then the last substituted fragment is eliminated from consideration, and the atoms forming the identified fragment having the next highest context dependent probability for one of the partial coordinate string is substituted into its corresponding partial coordinate string, thereby to define a second updated partial coordinate string. This updated partial coordinate string is tested in accordance with the predetermined test parameters. The substitution is repeated until the substitution of a database fragment meeting the criterion is found for substantially all of the partial coordinate strings.

The invention is preferably implemented on an appropriately programmed digital computer.

The coordinates of the atoms in a partial coordinate string may be derived from a variety of experimental or theoretical sources. Partial coordinate string sets may result from low resolution x-ray structural studies, Nuclear Magnetic Resonance (NMR), Nuclear Overhauser Effect (NOE) (or similarly obtained) sets of interatomic distances, structural templates from homologous proteins, or theoretically derived models for protein structure.

The present invention achieves its high level of performance in generating accurate structures through a hierarchical procedure of evaluation of polypeptide conformations that are acceptable in a given structural context. The ranking of alternative acceptable conformations is derived in part from a statistical survey of protein structures known in

detail from x-ray crystallographic studies, together with other steric and energetic considerations.

The present invention may be used in any of a variety of applications. For example, the present invention may be used to model proteins from alphaC coordinates, such as those obtained from initial x-ray electron density maps or from NMR techniques. The invention may be used for protein homology modeling or in protein engineering applications. Exemplary of the latter are the determination of amino acid sequences that are consistent with the formation of a stable protein structure, determination of stable structures with linked substitutions, or the design of integral fusion proteins, which are proteins linked together in a specified geometrical relationship.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be more fully understood from the following detailed description, taken in connection with the accompanying drawings, which form a part of this application and in which:

Figure 1 is a stylized pictorial representation of hypothetical molecular structure used to illustrate the principles of the present invention whereby the three-dimensional conformation of the molecule may be derived;

Figure 2 is a flow diagram of the steps of the method in accordance with the present invention.

DETAILED DESCRIPTION OF THE INVENTION

Throughout the following detailed description, similar reference numerals refer to similar elements in all Figures of the drawings.

Figure 1 is a a stylized pictorial representation of hypothetical molecular structure used to illustrate the principles of the present invention while Figure 2 illustrates in flow diagram form the steps of a method in accordance with the present invention whereby the substantially complete set of three-dimensional coordinates of the molecule may be constructed from a partial coordinate set.

The molecule under investigation, generally indicated by the character M, will in most cases be a protein molecule, but it should be understood that the present invention may be used to predict the three-dimensional configuration of any molecular structure. As suggested in Figure 1 the molecule M in most cases presents a complex three-dimensional shape. The molecule M will generally have a plurality of atoms, schematically indicated by the character A, that are bonded together to define the main structural framework of the molecule M. In the case of a protein, this framework is known as the backbone. In the most general case, one or more sidechain(s) of atoms, indicated by the character C, extend from the main structural framework of the molecule M. In Figure 1 the sidechain C is shown as extending from a helical portion of the molecule M. The atoms A forming the sidechains C may occupy certain angular positions (rotomer positions) with respect to the main structural framework of the molecule. In addition other molecular structures M' may be incorporated into or bound to the molecule M.

The present invention relates to the determination of the three-dimensional shape (i.e., conformation) of a molecule, as

the molecule M, based upon information relative to the locations of certain of the atoms A known to reside in some predetermined sequence at some predetermined region of the molecule M. In the preferred case the invention is implemented on a serial computer such as that manufactured by Digital Equipment Corporation and sold as a VAX.

The first step of the invention, as generally indicated by the block 10, is the definition of at least two or more partial coordinate strings. Each partial coordinate string (PCS) includes the three-dimensional coordinates of the location(s) of one or more atoms known to reside at some predetermined region (and in some predetermined sequence, for the case of plural atoms) of the molecule. The string may, for example, represent atoms that comprise the alpha-carbon atom (alphaC) backbone of the molecule (e.g., atoms similar to the atoms A shown in Figure 1), may represent side chains extending from the backbone (e.g., atoms similar to those in the sidechains C shown in Figure 1), or other atoms incorporated into, or bound to, a protein molecule (e.g., atoms similar to those in the molecule M' shown in Figure 1).

To define a partial coordinate string the coordinates of the locations of known atoms are first expressed with reference to any convenient coordinate system. This activity is illustrated by the block 10. Typically the coordinate system used would be a Cartesian system, as is illustrated in Figure 1. However, in some instances representations of atomic locations with reference to a fixed system may not be possible. Instead,

the coordinates of atomic locations may b expressed in terms of their relationship with each other.

The coordinates might be obtained in any of a variety of ways including: (1) from a low resolution X-ray structure determination; (2) from a model whose structure is anticipated (based, for example, on amino acid sequence homology) to be similar to the protein whose structure is to be reconstructed; (3) from the results of nuclear magnetic resonance (NMR) measurements; (4) from computer graphics molecular modeling; or (5) from theoretical means.

Once the coordinates of the known atoms are assigned, sets of atoms are grouped together thereby to define the partial coordinate string. This activity is illustrated by the block 10. The partial coordinate strings may be assigned either randomly or based upon any additional information that may be available (and typically supplied as input data to the program). For example, reconstruction of a protein model from alphaC backbone coordinates assigns amino acid sequence positions in the polymeric polypeptide chain to the partial coordinate strings.

Any plural number up to Q partial coordinate strings may be defined where Q is at least two. The upper limit on the number of partial coordinate strings is based on the number of atoms in the molecule (e.g., this could range from several thousand to several hundred thousands). Each grouping of atoms whose coordinates form each partial coordinate string may have some or none of its members in common with other groupings. In most instances, however, each string will have at least one member in common. Moreover, the strings will typically be of the same predetermined length (e.g., six to ten alphaC atoms in length). It should be understood that it is only the locations of atoms known or believed to be present along

the molecule that are included within a partial coordinate string. Accordingly, the concept of a partial coordinate string is also meant to encompass voids, that is, locations where the actual presence or absence of an atom is unknown.

As is suggested in Figure 1, each partial coordinate string may be envisioned as a "window" W of predetermined length that sweeps along the molecule M. The window W is such that at any given instant it may contains a predetermined number of the atoms (or voids) of the molecule. The group of atoms disposed within the window W at any given instant defines the partial coordinate string. Successive partial coordinate strings may contain at least one atom in common, as illustrated at reference character X in Figure 1. It should be understood that it lies within the contemplation of this invention to alter partial coordinate string assignments at successive stages during the implementation of the invention. For example, an alphaC coordinate set running from 1 to n, with partial coordinate string assigned in increments of six, (using the polymer sequence numbering), can be redefined into new groups starting with coordinate n+1, for the purpose of searching for alternative database fragments that better or alternatively correspond to the partial input coordinate set.

As illustrated at the block 20 a structural context mask (SCM) is next defined for each coordinate string. The structural context mask is then used to search (as shown at block 30) in the manner of a template in a database, or library, search to be described. The library contains a predetermined number of molecular structures, each of which is subdivisible into many structural fragments (referred to as "DBF" -- database fragment-- in Figure 2). Each fragment is, in essence, a set of coordinate locations (in the predetermined coordinate system) for each atom disposed within each molecular fragment. As will be developed the structural context mask is used to define

a predetermined range of similarity about each partial coordinate string in order that a determination can be made as to determine whether a given structural fragment may be classified as similar, and thus a possible correspondence to, a given partial coordinate string. The search thus identifies each structural fragment that has at least some predetermined subset of the atoms thereof with coordinates that correspond within the predetermined limits imposed by the predetermined structural context mask to the coordinates of the atoms in each partial coordinate string.

In the case of protein modeling, an appropriate library is a subset of the known three-dimensional protein structures, such as those obtained from high resolution x-ray crystallography and contained in the Brookhaven Protein Data Bank. This library of molecular fragments is discussed in Bernstein et al., "The Protein Data Bank: A Computer-Based Archival File For Macromolecular Structures", (1977) J. Mol. Biol., 112, 535-542.

The database search for fragments that correspond to each partial coordinate string is made more efficient and effective by application of the structural context mask. A given structural context mask may be defined in various possible ways. For example, a suitable structural context mask may take the form of an root-mean-square spatial superposition error between potentially corresponding atoms of a fragment and of a partial coordinate string. In more complex applications a suitable structural context mask may take the form of correspondence of three-dimensional fit plus any specific features of molecular structure, amino acid sequence, or environment between such potentially corresponding atoms.

A structural context mask may also specify particular recurrent structural features in proteins. The structural context

mask may thus be based on contiguous or non-contiguous sets of residues in the structure. An example of the former would be a structural context mask representing a hairpin 1 op structure with Glycine-Proline sequence at the central positions in the loop where it reverses direction. An example of the latter might be a pair of antiparallel beta sheet strands connected by an interchain disulfide bond formed between the side chains of two cysteine residues.

Partial coordinate strings to be fit by fragments can correspond to covalently contiguous or noncontiguous sets of amino acid residues. This feature makes it possible to evaluate a partial coordinate string for, and reconstruct regions of, structure from sets of partial coordinate strings that have a defined three-dimensional relationship in a structure, but whose detailed geometry of covalent backbone connectivity is ambiguous or unknown.

To perform the search the structural context mask is preferably applied sequentially. In the protein modeling case, there may, for example, be performed an initial sort through the library of fragments to find those that have some degree of amino acid sequence similarity to the partial coordinate string under consideration. After fragments have been retrieved that meet some criteria of sequence correspondence subsequent tests may be applied based on structural correspondences between the fragments and the partial coordinate string. Because the database search for structural correspondences is a computationally intensive operation, use is made of the method described by Jones and Thirup, "Using Known Substructures In Protein Model Building and Crystallography", (1986) Embo. J., 5, 819-822. This article illustrates that fragments derived from a protein structural database can be used in an interactive computer graphics program to aid in fitting protein molecular models to crystalloraphically produced electron density maps.

In accordance with this technique coordinates are stored in interatomic distance matrices, thereby to sort and compare efficiently database coordinates for comparison with partial coordinate strings.

In typical operation each structural context mask is dynamically adjusted. This adjustment is indicated diagrammatically in Figure 2 by the line 42 extending from the block 40 to the block 20. For example, when partial coordinate strings comprising the locations of backbone atoms representing successive amino acid residue positions in a protein structure are to be reconstructed as complete polypeptide sidechains, each partial coordinate string would initially be fit to a structural fragment using a structural context mask spanning nine successive backbone alphaCarbon positions. The specific amino acid types (e.g., Histidine or Alanine) used are left unrestricted except for the central one, which is required to be an exact match in type to the corresponding amino acid in the structure. Typically the database of refined protein structures will be searched for a minimum of thirty fragments that meet an structural context mask criterion that the root mean square superposition error between the alphaCarbon atoms of the partial coordinate string and the fragment be less than 0.5 Angstroms. If thirty fragments that meet the structural context mask are not found, then the length of the partial coordinate string is reduced alternately at either end by one residue position, to a minimum partial coordinate string length of three residues.

Alternatively, in some situations, it is possible to alter the structural context mask either with respect to the number of fragments retrieved (e. g., reduce the number to twenty), alter the root mean square superposition criterion (e.g., increase root mean square superposition error to 1.0 Angstroms), or the

specification of amino acid residue type corresponding to specific positions in the partial coordinate string.

The decision about which set of a fragment's coordinates (or subfeature of the retrieved fragment's properties) is the best substitute for each partial coordinate string is based on computation of a context dependent probability (CDP). The context dependent probability is a measure of the "substitutability", or the degree of possible correspondence between a fragment identified in the search and a given partial coordinate string. This step in the method of the present invention is indicated at block 40.

The context dependent probability can be also be defined in a variety of ways. In the most simple instance a context dependent probability can be defined as a root-mean-square fit between fragment and partial coordinate string, or it can reflect the statistics of structural patterns in the context of a complex structural context mask. In some cases, a fragment's context dependent probability may be operationally based on previous experience. For example, context dependent probability thresholds may be set that determine when a fragment is accepted as a substitute for a partial coordinate string during a given stage of model building.

As indicated by block 45 when all, or substantially all, of the partial coordinate strings have been associated with fragments identified by the search, and the corresponding context dependent probability of each fragment evaluated, all of the fragments identified by the search are ranked in order of context dependent probability. Rank values are assigned numerically according the degree of correspondence between the partial coordinate string and the retrieved fragment using the criteria defined by the context dependent probability. For example, when partial coordinate strings comprising the

locations of alphaC atoms representing successive amino acid residue positions in a protein structure are to be reconstructed as complete polypeptide chains, each partial coordinate string is ranked according to its root mean square superposition error with corresponding atoms of the partial coordinate string.

In more complicated situations, such as when amino acid residue sidechain groups are appended to a backbone polypeptide chain, then the observed distributions that are found generally represent the discrete sampling of alternative sidechain rotomer conformational states. In this case, statistics are compiled on the relative frequencies of occurrence of these discrete conformational states in the retrieved fragments for each partial coordinate string, in the appropriate local context of the associated structural context mask. The statistics obtained are then used to assign numerical context dependent probabilities to the alternative rotomer states for each partial coordinate string. For the backbone reconstruction process, the context dependent probabilities, as measured by the best partial coordinate string-fragment root mean square fit are used to numerically rank partial coordinate strings in the order that they are to be substituted. For example, structural context masks can be defined that correspond to new or alternative patterns of covalent, or noncovalent, linkages between contiguous or noncontiguous parts of an existing or novel protein structure. Applications include the engineered introduction of new linkages or binding sites into protein backbones, or the determination of structures using NMR distance data.

As indicated at block 50 the full coordinate set of the atoms contained in the fragment having the highest context dependent probability is substituted into its corresponding partial coordinate string, thereby to define an updated, more

complete coordinate set for that coordinate string. This operation is indicated at block 52. After each substitution of a partial coordinate string by a fragment, the updated partial coordinate string that results from the substitution (now composed of both original partial coordinate strings and any substituted fragments) is tested against the partially reconstructed remainder of the molecule using steric (e. g., van der Waals exclusion) considerations. Additionally, if desired, computational techniques such as energetic analysis, statistical correlations or other predetermined criteria of acceptability may be used.

The results of the test block 54 form the basis of a decision block 56. If the steric and energetic test results are acceptable according to the predefined acceptance criteria, the substitution of the block 52 is repeated, using the fragment in the pool of fragments identified by the search having the next highest context dependent probability and substituting the same into its corresponding partial coordinate string. Repeating the substitution action of the block 52 thereby define a second, updated, more complete coordinate set for the coordinate string. The second updated coordinate string thus includes original (i.e., as yet unsubstituted) portions of the partial coordinate string, and the coordinate sets corresponding to the substituted fragments. The second updated partial coordinate string is subjected to the test of the block 56, and assuming passage of the test, the substitution looping continues. In practice, the reiteration of the activities in the block 52 is dependent upon the presence of other partial coordinate strings for which substitutions have not yet been made. This condition is contemplated by the presence of the decision block 58. Thus, substitution of the remaining partial coordinate strings continues to completion under control of the block 58, or, as will be discussed, until an updated partial coordinate string fails the test in the block 56.

The underlying principle of the present inv ntion is to define a strategy for rebuilding structures from partial coordinate strings that orders the pattern of fragment substitution so that those fragments that represent three dimensional structural patterns of highest probability are substituted first. The process of substituting fragments in this manner will, owing to steric interference or other tests applied at each stage of the reconstruction process, excludes from further consideration some fragments with lower (more indeterminate) context dependent probabilities. In contrast, substitution of fragment in a random or arbitrary sequence (e. g., 1 to n) often results in the incorrect placement of a single residue creating a "domino-effect" cascade of errors in placement of subsequent fragments.

If an updated partial coordinate string fails to meet the predetermined criteria of the test in in the block 54, as indicated by the "No" branch from the decision block 56, the fragment having the last previously substituted fragment is deleted from the updated partial coordinate string. The coordinate set of the fragment having the next highest context dependent probability is then substituted into the partial coordinate string for which if has been identified as a possible correspondent. This action is indicated in the block 60.

If all partial coordinate strings could not be built in the context of the starting data, as indicated by the "No" branch from the block 64, the starting data may be partially incorrect. Analysis of the failure patterns usually suggest which regions of the starting data are at fault. For example, a large number of sidechain rebuilding failures in a given region of structure may suggest that the corresponding local section of backbone is reconstructed incorrectly. Alternatively, the structure produced may be analyzed for a variety of characteristics

related to stability or functional properties. These include the application of computational energy methods to estimate the contributions of Van der Waals, electrostatic and conformational interactions to the protein stability, computations to estimate the structures packing density, extent and type of residues interacting with surrounding solvent, and molecular dynamics simulations to estimate entropy contributions to stability or determined location of strained regions in the structure.

However, if substantially all of the partial coordinate strings have been substituted with fragments identified by the search the resulting updated partial coordinate strings meet the tests of the block 54, the resulting final updated partial coordinate string defines the set of three-dimensional coordinates for the molecule of interest. The structures so produced may be analyzed, as indicated at the block 68 for a variety of characteristics related to stability or functional properties. The tests applied are similar to those discussed immediately above.

The invention as diagrammatically shown in Figure 2 can itself be used as a structural analysis program, since it provides a measure of how well any structure corresponds to an experimentally determined structural database. The structure to be analyzed may be a structure determined experimentally, or be the result of a combination of experimental data and the reconstruction operations in accordance with the present invention. Moreover, even complete models generated using the method of the present invention from theoretically derived framework structures (e.g. cylindrical sheets or curved beta sheets or bundles of alpha-helices) can be usefully analyzed by using different structural context masks to estimate context dependent probabilities in the initial structure generation, and subsequent, structure analysis stages. Additionally, regio..

that have failed successful reconstruction under the criterion of an initial structural context mask, can be both analyzed and reconstructed using an alternative (and possibly more specialized) structural context mask appropriate to the specific partial coordinate string which as incorrectly reconstructed during an initial pass of the program.

The logical steps outlined in Figure 2 can be utilized with several different types of input (e. g., different types of partial coordinate strings), utilized with different structural context masks, utilized with different methods of evaluating context dependent probabilities for retrieved fragments, and produce correspondingly different outputs, which essentially represent hierarchical stages in protein structure organization. Accordingly, utilization of the teachings of the present invention in various applications will generally incorporate a hierarchical approach to structure reconstruction, with different levels of structure being built up and tested in successive stages of application of the program. The different applications thus differ according to the sequence and specificity with which structural features are introduced and tested during the process of generating a new model protein structure.

As mentioned earlier the present invention may be implemented on a general purpose digital computer. It should be understood, therefore, that when such a general purpose digital computer is operated in accordance with a program such a programmed machine defines an apparatus for predicting the three-dimensional configuration of a molecule in accordance with this invention.

Those skilled in the art, having the benefit of the teachings of the present invention, may impart numerous modifications thereto. It should be understood, however, that such modifications are to be construed to lie within the contemplation of the present invention, as defined by the appended claims.

WHAT IS CLAIMED IS:

1. A method of predicting the three-dimensional configuration of a molecule comprising the steps of:

(a) defining, in accordance with a predetermined coordinate system, at least a first and a second partial coordinate string, each partial coordinate string comprising the three-dimensional coordinates of the location of each of a predetermined number of atoms known to reside in some predetermined sequence at some predetermined region of the molecule;

(b) searching a library of known molecular structures to identify each structural fragment that has at least some predetermined subset of the atoms thereof with coordinates that correspond within limits imposed by a predetermined structural context mask to the coordinates of the atoms in the each partial coordinate string;

(c) assigning to each structural fragment identified by the search a context dependent probability that the particular structural fragment corresponds to a predetermined partial coordinate string;

(d) substituting the coordinates of the atoms forming the structural fragment having the highest context dependent probability into its corresponding partial coordinate string, thereby to define an updated partial coordinate string; and

(e) testing the resultant updated partial coordinate string in accordance with predetermined test parameters.

2. The method of claim 1 wherein there are at least two atoms in each partial coordinate string.

3. The method of claim 2 wherein each partial coordinate string has at least one atom in common.

4. The method of claim 1 wherein a different predetermined structural context mask is defined for each partial coordinate string.

5. The method of claim 1 further comprising the steps of:

(f) based upon the results of the test substituting into the partial coordinate string the coordinates of the atoms forming the structural fragment having the second highest context dependent probability into its corresponding partial coordinate string, thereby to define a second updated partial coordinate string; and

(g) testing the resultant second updated partial coordinate string in accordance with the predetermined test parameters.

6. The method of claim 5 further comprising the steps of:

(h) repeating steps (f) and (g) until substitutions have been made into substantially all partial coordinate strings.

7. The method of claim 1 further comprising the steps of:

(f) based upon the results of the test eliminating and discarding the last substituted fragment from the updated partial coordinate string, and substituting

into the partial coordinate string the coordinates of the atoms forming the structural fragment having the second highest context dependent probability into its corresponding partial coordinate string, thereby to define a second updated partial coordinate string; and

(g) testing the resultant second updated partial coordinate string in accordance with the predetermined test parameters.

8. The method of claim 7 further comprising the steps of:

(h) repeating steps (f) and (g) until substitutions have been made into substantially all partial coordinate strings.

9. The method of claim 1 wherein the molecule is a protein molecule and wherein at least one partial coordinate string includes the three-dimensional coordinates of the location of some predetermined number of atoms known to reside in a side chain located at some predetermined position along the protein molecule.

10. The method of claim 1 wherein the molecule is a protein molecule having a backbone defined therein, and wherein at least one partial coordinate string includes the three-dimensional coordinates of the location of some predetermined number of atoms known to reside in a side chain located at some predetermined position along the backbone of the protein molecule.

11. A method of predicting the three-dimensional configuration of a molecule comprising the steps of:

(a) defining, in accordance with a predetermined coordinate system, a plurality of partial coordinate strings, ach partial coordinate string comprising the three-dimensional coordinates of the location of each of a predetermined number of atoms known to reside in some predetermined sequence at some predetermined region of the molecule;

(b) defining a structural context mask for each partial coordinate string;

(c) searching a library of known molecular structures to identify each structural fragment therein that has at least some predetermined subset of the atoms thereof with coordinates that correspond within limits imposed by a predetermined structural context mask to the coordinates of the atoms in the partial coordinate string corresponding to that structural context mask;

(d) assigning to each structural fragment identified by the search a context dependent probability that the particular structural fragment may be substituted for the predetermined partial coordinate string;

(e) selecting from all of the structural fragments identified by the search the fragment having the highest context dependent probability,

(f) substituting the coordinates of the atoms forming that fragment into its corresponding partial coordinate string, thereby to define an updated partial coordinate string;

(g) testing the resultant updated partial coordinate string in accordance with predetermin d test parameters and eliminating and discarding the last substituted fragment if the updated partial coordinate string fails to meet a predetermined acceptance criterion;

(h) repeating steps (e), (f) and (g) until substitutions have been made for substantially all of the partial coordinate strings.

12. The method of claim 11 wherein wherein the structural context mask for a selected partial coordinate string is changed based upon the number of fragments identified by the search as corresponding to that partial coordinate string.

13. Apparatus for predicting the three-dimensional configuration of a molecule comprising:

(a) means for defining at least a first and a second partial coordinate string, each partial coordinate string comprising the three-dimensional coordinates of the location of each of a predetermined number of atoms known to reside in some predetermined sequence at some predetermined region of the molecule;

(b) means for searching a library of known molecular structures to identify each structural fragment that has at least some predetermined subset of the atoms thereof with coordinates that correspond within limits imposed by a predetermined structural context mask to the coordinates of the atoms in the each partial coordinate string;

(c) means for assigning to each structural fragment identified by the search a context dependent probability that the particular structural fragment corresponds to a predetermined partial coordinate string;

(d) means for substituting the coordinates of the atoms forming the structural fragment having the highest context dependent probability into its corresponding partial coordinate string, thereby to define an updated partial coordinate string; and

(e) means for testing the resultant updated partial coordinate string in accordance with predetermined test parameters.

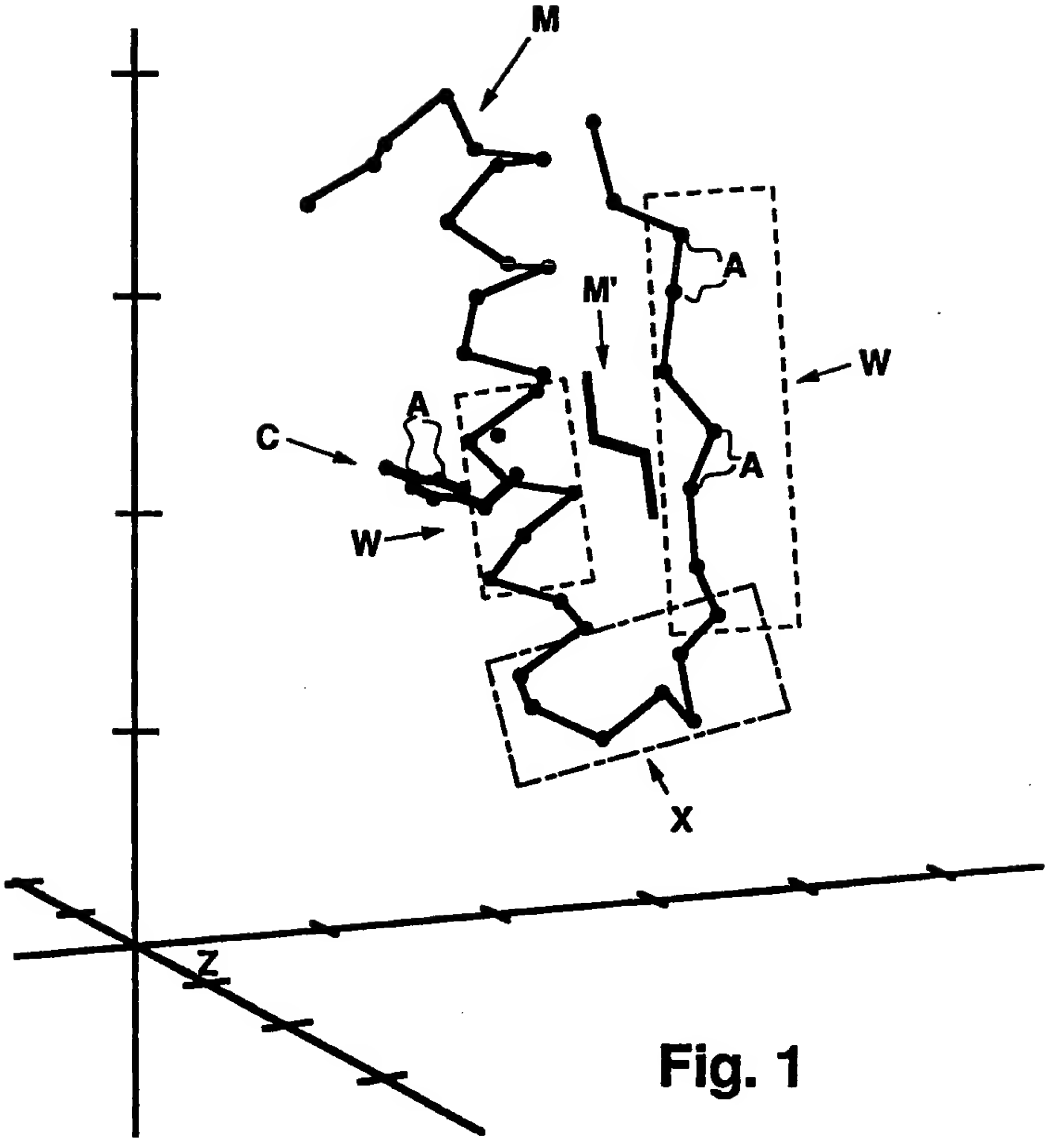


Fig. 1

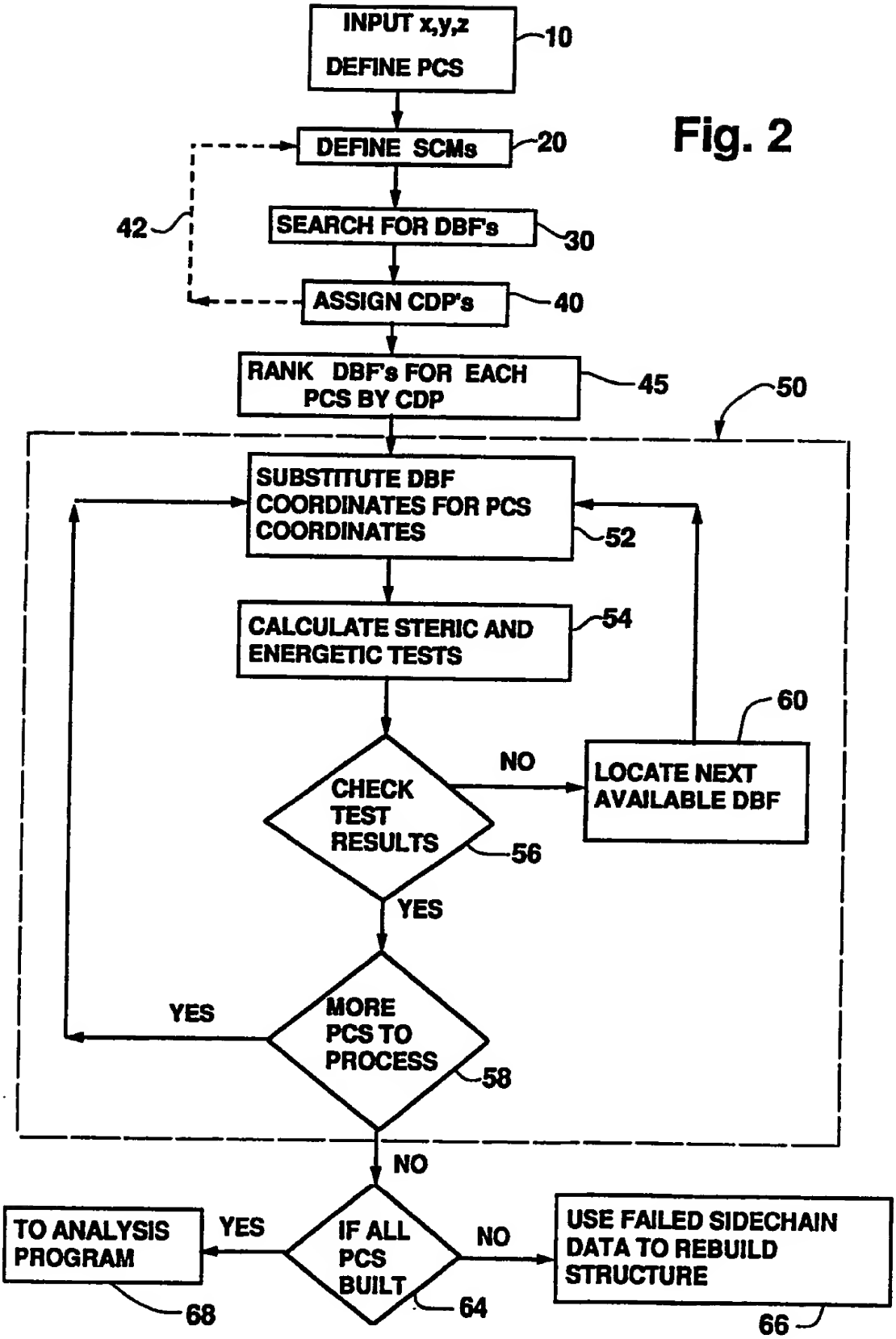

INPUT x,y,z
DEFINE PCS
10
DEFINE SCMs
20
42
SEARCH FOR DBF's
30
ASSIGN CDP's
40
RANK DBF's FOR EACH PCS BY CDP
45
50
SUBSTITUTE DBF COORDINATES FOR PCS COORDINATES
52
CALCULATE STERIC AND ENERGETIC TESTS
54
60
CHECK TEST RESULTS
56
NO
LOCATE NEXT AVAILABLE DBF
YES
MORE PCS TO PROCESS
58
YES
NO
IF ALL PCS BUILT
64
YES
TO ANALYSIS PROGRAM
68
NO
USE FAILED SIDECHAIN DATA TO REBUILD STRUCTURE
66

Fig. 2

INTERNATIONAL SEARCH REPORT

International Application No PCT/US91/01106

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all)

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC(5): G01N 33/00
US CL.: 364/496,499; 436/183

II FIELDS SEARCHED

Minimum Documentation Searched	
Classification System	Classification Symbols
US	364/496,499 436/183

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched

III. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
A	US, A, 4,511,841 (BARTUSKA et al.) 16 April 1985 See column 1, lines 11-25.	1-13
A	US, A, 4855,931 (SAUNDERS) 08 August 1989 See abstract and figs. 5-8 & 12.	1-13

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive st[illegible]hen the document is combined with one or more other documents, such combination being obvious to a pe[illegible] skilled in the art.
"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
10 APRIL 1991	29 APR 1991
International Searching Authority	Signature of Authorized Officer
ISA/US	EDWARD R. COSIMANO

[illegible] GOC-HO
INTERNATIONAL DIVISION

Form PCT/ISA/210 (second sheet) (May 1986)

FURTHER INFORMATION CONTINUED FROM THE SECOND SHEET

V ☒ OBSERVATIONS WHERE CERTAIN CLAIMS WERE FOUND UNSEARCHABLE [1]

This international search report has not been established in respect of certain claims under Article 17(2) (a) for the following reasons:

1. ☒ Claim numbers 1-13, because they relate to subject matter not required to be searched by this Authority, namely:

they relate to mathematical theories under Article 17(2)(a)(i). Claim 13 is considered to be a method claim, since a specfic device(s) are not claimed.

2. ☐ Claim numbers , because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out [2], specifically:

3. ☐ Claim numbers_____, because they are dependent claims not drafted in accordance with the second and third sentences of PCT Rule 6.4(a).

VI ☐ OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING [3]

This International Searching Authority found multiple inventions in this international application as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims of the international application.

2. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims of the international application for which fees were paid, specifically claims:

3. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claim numbers:

4. ☐ As all searchable claims could be searched without effort justifying an additional fee, the International Searching Authority did not invite payment of any additional fee.

Remark on Protest

☐ The additional search fees were accompanied by applicant's protest.

☐ No protest accompanied the payment of additional search fees.